LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-416455

Feature Extraction and Selection From the Perspective of Explosive Detection

S. K. Sengupta

September 2, 2009

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Feature Extraction and Selection
From the Perspective of Explosive Detection

Sailes K. Sengupta

LLNL/NDE/EE
LLNL-TR-416455

March 1998

1. Introduction

Features are extractable measurements [1-3] from a sample image summarizing the information content in an image and in the process providing an essential tool in image understanding. In particular, they are useful for image classification into pre-defined classes or grouping a set of image samples (also called clustering) into clusters with similar within-cluster characteristics as defined by such features. At the lowest level, features may be the intensity levels of a pixel in an image. The intensity levels of the pixels in an image may be derived from a variety of sources. For example, it can be the temperature measurement (using an infra-red camera) of the area representing the pixel or the X-ray attenuation in a given volume element of a 3-d image or it may even represent the dielectric differential in a given volume element obtained from an MIR image. At a higher level, geometric descriptors of objects of interest in a scene may also be considered as features in the image. Examples of such features are: area, perimeter, aspect ratio and other shape features, or topological features like the number of connected components, the Euler number (the number of connected components less the number of 'holes'), etc. Occupying an intermediate level in the feature hierarchy are texture features which are typically derived from a group of pixels often in a suitably defined neighborhood of a pixel. These texture features are useful not only in classification but also in the segmentation of an image into different objects/regions of interest.

At the present state of our investigation, we are engaged in the task of finding a set of features associated with an object under inspection (typically a piece of luggage or a brief case) that will enable us to detect and characterize an explosive inside, when present. Our tool of inspection is an X-Ray device with provisions for computed tomography (CT) that generate one or more (depending on the number of energy levels used) digitized 3-dimensional attenuation images with a voxel resolution of the order of one quarter of a milimeter. In the task of feature extraction and subsequent selection of an appropriate subset thereof, several important factors need to be considered. Foremost among them are:

1. Definition of the *sampling unit* from which the features will be extracted for the purpose of detection/ identification of the explosives.

2. The *choice of features* (given the sampling unit) to be extracted that can be used to signal the existence / identity of the explosive.

3. *Robustness* of the computed features under different inspection conditions.

To attain robustness, invariance under the transformations of translation, scaling, rotation and change of orientation is highly desirable.

4. The *computational costs* in the process of feature extraction, selection and their use in explosive detection/ identification

In the search for extractable features, we have done a thorough literature survey with the above factors in mind and come out with a list of features that could possibly help us in meeting our objective. We are assuming that features will be based on *sampling units that are single CT slices of the target*. This may however change when appropriate modifications should be made to the feature extraction process. We indicate below some of the major types of features in 2- or 3-dimensional images that have been used in the literature on application of pattern recognition (PR) techniques in image understanding [3, 4] and are possibly pertinent to our study. In the following paragraph, we briefly indicate the motivation that guided us in the choice of these features, and identify the nature of the constraints.

The principal feature types derivable from an image will be discussed in section 2. Once the features are extracted, one must select a subset of this feature set that will retain the most useful information and remove any redundant and irrelevant information that may have a detrimental effect on the classifier performance. This is discussed in section 3. Section 4 provides a brief summary.

2. Feature Types.

Depending on their origin, features can be broadly divided into the following categories:

(a) Features in the spatial domain
(b) Textural Features
(c) Features in the transform domain
(d) Shape Features

Of these, the shape features are unlikely to play a significant role in the specific explosive detection task we will be involved with. Here we are trying to detect/identify explosive materials which are more or less *homogeneous and do not have any defined shape*. The shape of the container is irrelevant for our purpose although this may change in any specific circumstance. We shall briefly elaborate the other three in the following.

Spatial domain features.

These are perhaps the most well understood and widely used features to be found in the pattern recognition (PR) literature. These include (i) the amplitude (or spectral) features and the (ii) histogram features.

The amplitude features include entities like, reflectivity, IR temperatures obtained from a hyperspectral camera, RGB color components, X-ray attenuation at different energy levels, the dielectric differential measured by a microwave impulse radar etc. These refer
directly to the physical domain from which they were created. Histogram features on the other hand are derived from the frequency distribution of the amplitudes mentioned above. Typically, they include quantities like the statistical moments, the entropy and other statistical parameters associated with the distribution [3]. The histogram features and the invariant moments are listed in Appendix 1.

In the context of statistical moments it may be important to consider what is known as *invariant moments* such as Zernike moments [4, 5] which are invariant under rotations, scale transformations and translations (RST). Fast computation algorithms for Zernike moments are presented in [4] and [5] indicates that these may be made 'illumination invariant' as well.

Textural Features. Texture is often interpreted in the literature as a set of statistical measures of the *spatial* distribution of gray levels in an image. Here it is assumed that texture information is contained in the average spatial relationships that gray levels have with one another [6, 7]. Within this category there are several sub-categories that have been found useful. These are based on: Gray Level Co-occurrence Matrix (GLCM), Sum and Difference Histogram (SADH) and , Gray Level Difference Vector (GLDV). In the following we briefly describe each of these [6-11].

GLCM. The gray-level co-occurrence matrix method assumes that textural information is characterized by a set of co-occurrence matrices P_d (i, j) where the (i, j) th element is the relative frequency with which two pixels with gray values i and j respectively and separated at a vector distance d, occur in the image. [6, 7]

SADH. Unser [10] proposed the sum and difference histogram method in which the second order probability function of a co-occurrence matrix is replaced by estimates of the first order probability functions along the principal axes of the co-occurrence matrix. Chen et al. [11] showed that the SADH method produced classification accuracies equivalent to those obtained using the GLCM method but with significant savings in computing resources. Appendix 2 provides the computational formulas for these features.

GLDV. The gray level difference vector [11] approach is based on the absolute differences between pairs of gray levels I and J found at a vector distance d apart. The difference-vector probability density function P_d (m) is defined for m = | I - J | , where I and J are the corresponding gray levels, and is obtained by normalizing the gray-level frequencies of occurrence by the total frequencies. These textural measures are computed in Appendix 3.

Features in the transform domain.

Frequency domain information in an image is contained in image transforms such as Fourier or Gabor. Fourier transform features are selected by computing the total power spectrum in annular rings, wedge-shaped sectors or strips (horizontal or vertical). Of these, the first measures the coarseness (fineness) of the texture and the second measures the angular sensitivity [6]. Other power spectrum measures have been defined in [11, 12]. These features, particularly the last two appear to be of limited value in the context of explosive detection since explosives or simulants are unlikely to possess texture that is periodic or directional in nature.

Extensions to X-Ray CT Images.

Much of the above discussion extends in a straightforward manner to the case of 3-dimensional images such as digitized CT images although in this case, the computational complexity is enhanced by at least an order of magnitude. It should be noted that special care must be taken in using the textural features indicated above where reference to a vector difference has been made. Typically, for the sake of simplicity, an implicit assumption of isotropy is made in textural feature computation of 2-d images. This is much less likely to be valid for CT images in general because the vertical resolution in a CT image is frequently not the same as that in the two orthogonal horizontal directions. However this problem can be averted when our sampling unit is a slice or an extracted region from a slice rather than an entire volume or in the case of a 3-dimensional sampling unit, the spatial resolutions along the three axes are the same. In all other CT images, the features that use vector separation of pixels, the question of isotropy must be addressed. Features derived in the transform domain are extended readily to the 3-d case albeit at the expense of additional computational complexity.

3. Feature subset selection

It is well known from statistical decision theory that probability of classification error decreases when additional measurements (features) are taken into consideration. This is true only for infinite sample sets for which the estimation errors for the system parameters can be ignored. In practice however, only finite training sets are available for the purpose of supervised clsssification and the estimation errors are no longer negligible. Due to these errors, the system can be so finely tuned to the training set that it lacks generalization capability. Since the number of parameters and the associated estimation errors increase rapidly with dimension, it may be advantageous to sacrifice some useful information in order to keep the number of these parameters to a minimum.

The criterion function used to assess the discriminatory power of the individual feature set is the overall probabilistic distance measure between classes to be discriminated. This measure uses the complete information about the probabilistic structure of classes in a classification procedure. This information is given in terms of the class conditional probability density functions and the prior class probabilities. One such measure of class separability that is frequently applied in practice is based on the overall Bhattacharya distance between several multivariate populations [13]. This measure C_b is given by the expression:

$$C_b = \sum_{i=1}^{N} \sum_{j=i+1}^{N} J_B(i,j),$$

where N is the number of classes and $J_B(i,j)$ is the Bhattacharya distance between two multivariate populations i and j given by

$$J_B(i,j) = \left(\mu_i - \mu_j\right)^t \Sigma^{-1} \left(\mu_i - \mu_j\right).$$

Here, μ_i and μ_j are the feature mean vectors of the two populations and Σ the (assumed common) covariance matrix of the features for these populations. The prior probabilities of the classes have been assumed to be equal above.

Several algorithms exist in the literature for feature subset selection preserving their class discriminatory power. Among them there is one that is commonly used is called the sequential forward selection (SFS) method. This is suboptimal but efficient and generally produces good results[12]. The algorithm is as follows:

> Let Y be the set of D features $Y = \{Y_i, i=1,2,...,D\}$. Suppose at the k th step, k features have been selected from feature set X_k. Rank the elements F_i of the set of available features $Y - X_k$ so that
>
> $$C_b(X_k + F_1) \geq C_b(X_k + F_2) \geq ... \geq C_b(X_k + F_{(D-k)})$$
>
> Then the feature set $X_{k+1} = X_k + F_1$ is created at the (k+1) th step. The feature inclusion process is continued until the required number of features is chosen. The algorithm is initialized by setting X_k to null.

The main drawback of SFS is that once a feature is included, it cannot be deleted, even though it may become redundant or irrelevant due to the features added later on. However, in spite of its suboptimality, it is an efficient way of selecting features useful for classification and of keeping the computational cost low.

4. Summary

A short review of the feature extraction and selection process with the perspective of CT data analysis has been presented. The features to be extracted are from the spatial or frequency domain with special emphasis on spatial domain features including textural features. Computational complexity may be a strong factor influencing the feature selection process. The need to pay attention to the possible lack of isotropy in textural features for the CT data has been pointed out. Finally, an efficient albeit suboptimal process of feature subset selection has been described.

References.

[1] Schalkoff R J , Digital Image Processing and Computer Vision, J. Wiley, Inc. New York, 1989.

[2] Fukunuga, K. , Introduction to Statistical Pattern Recognition, 2nd ed., Academic Press, San Diego, CA, 1990.

[3] Jain A K, Fundamentals of Digital Image Processing, Prentice Hall, Englewood Cliffs, NJ 1989.

[4] Belkasim S. O.; Ahmadi M. and Shridhar M., Efficient algorithm for fast computation of zernike moments. Journal of the Franklin Institute-engineering and applied mathematics, Jul 1996, V333B N4:577-581.

[5] Wang, L. Z. , Healey, G. Using Zernike moments for the illumination and geometry invariant classification of multispectral texture, IEEE Transactions on Image Processing, 1998 Feb., V7 N2:196 - 203.

[6] Levine M D, Vision in Man and Machine, McGraw Hill, Inc., New York, 1985

[7] Haralick R. M. , K. S. Shanmugan and D. I. Dinstein, Textural Features for Image Classification, IEEE Trans. Syst. Man. Cybern., SMC-3, 610-621.

[8] Welch, R. M., S. K. Sengupta, A.K. Goroch, P. Rabindra, N. Rangaraj and M. S. Navar, Polar Cloud and Surface Classification Using AVHRR Imagery: An Intercomparison of Methods, Jour. Appl. Meteor.,V. 31, No. 5, May 1992, pp. 405-420

[9] Welch, R. M., S. K. Sengupta and D. W. Chen, 1988: Cloud Field Classificaton Based Upon High Spatial Resolution Textural Features, Part I: Gray-level co-occurrence matrix approach. J. Geophys. Res., V. 93, 14 pp. 14767-14781.

[10] Unser, M. 1986: Sum and Difference Histograms for Texture Classification, IEEE Trans. Pattern Anal. Mach. Intell., PAMI-8, 118-125

[11] Chen, D. W., S. K. Sengupta, and R. M. Welch , 1989: Cloud Field Classification Based upon High Spatial Resolution Textural Features. Part II: Simplified Vector Approaches, Journal of Geophysical Research V. 94, 14749-14765.

[12] Garand, L., 1988: Automated Recognition of Oceanic Cloud Pattern, Part I: Methodology and Application to Cloud Climatology, J. Climate, V. 1, 20-39.

[13] Devijver, P. A. and J. Kittler, 1982: Pattern Recognition: A Statistical Approach, Prentice Hall, Englewood Cliffs, NJ 1982.

Appendix 1

Amplitude Features and Zernike Moments [Ref. 3-5]

Histogram Features:

Define the relative frequency $f(x)$ of the pixels with amplitude x in a given region as

$$f(x) = n(x) / N$$

where $n(x)$ is the number of pixels with amplitude x and N is the total number of pixels in the region. The following features are derived from this distribution and are commonly in use. The moments below are defined for $i = 1, 2, \ldots$. Note that for attenuation images we have nonnegative pixels only and hence the first and second measures below are the same.

The i th moment : $$m_i = \sum_{x=0}^{N-1} x^i f(x)$$

The i th absolute moment : $$m_{i,a} = \sum_{x=0}^{N-1} |x|^i f(x)$$

The i th central moment: $$\mu_i = \sum_{x=0}^{N-1} (x - m_1)^i f(x)$$

The i th absolute central moment: $$\mu_{i,a} = \sum_{x=0}^{N-1} |x - m_1|^i f(x)$$

Entropy: $$H = \sum_{x=0}^{N-1} f(x) \ln_2 f(x) \quad \text{bits}$$

Angular second moment: $$asm_1 = \sum_{x=0}^{N-1} [f(x)]^2$$

Zernike Moments.

Let f(x,y), (x,y) $\in$ D be an image with domain D. The Zernike moments of f are given by the projection of f on the Zernike polynomials V_{mn} :

(1) $$Z_{mn} = \frac{n+1}{\pi} \iint f(x,y) V_{mn}^*(x,y)\,dxdy$$ where the Zernike polynomials are (in polar coordinates) given by

(2) $$V_{mn}(\rho, \vartheta) = r_{mn}(\rho) \exp(jm\vartheta),\ j = \sqrt{-1}$$ and where

(3) $$r_{mn}(\rho) = \sum_{s=0}^{(m-|n|)/2} (-1)^s \frac{(m-s)!}{s!\left(\frac{m+|n|}{2} - s\right)!\left(\frac{m-|n|}{2} - s\right)!} \rho^{m-2s}$$

In the discrete case, (1) can be rewritten as

$$Z_{mn} = \frac{n+1}{\pi}\sum_{\rho}\sum_{\vartheta} f(\rho,\vartheta) r_{mn}(\rho)\exp(-jm\vartheta) \qquad \textbf{(4)}$$

where the sums are taken over the unit circle $\rho \le 1, 0 \le \vartheta \le 2\pi$. To compute the Zernike moments of an image $f(x,y)$ we map the largest circular subset of the domain of the function to the unit circle. The area outside the circle is not considered in the computation.

The Zernike moments are only rotation invariant. To obtain scale and translation invariance, the image is first subjected to a normalization process using its regular first and second moment. The rotation invariant features are then extracted from the normalized image.

Appendix 2

The Sum And Difference Histogram Features [Ref. 11]

From the gray level pair (I, J), two density functions $P_d^S(K)$ and $P_d^D(L)$ are defined that are respectively the histograms for the sum $K = I + J$ and the difference $L = I - J$. The histograms are normalized by dividing the individual frequencies by the sum of the component frequencies in each. The following provides the computational formulas for the SADH features. The subscripts (superscripts) S, D hasve been used to denote the sum and difference respectively. The subscript d denoting the vector separation of I and J is understood. The letters K, L have also been used as summation indexes.

Mean:

$$\mu_S = \sum_K P^S(K)$$

Standard deviation:

$$\sigma^2 = \left\{ \frac{1}{2} \left[\sum_K \left(K - \mu_S\right)^2 P^S(K) + \sum_L L^2 P^D(L) \right] \right\}^{1/2}$$

Note that

$$\sigma_S^2 = \sum_K \left(K - \mu_S\right)^2 P^S(K) \quad ,$$

$$\sigma_D^2 = \sum_K \left(K - \mu_D\right)^2 P^D(K)$$

so that

$$\sigma^2 = \left[\frac{1}{2} \left(\sigma_S^2 + \sigma_D^2 \right) \right]^{\frac{1}{2}}$$

Contrast:

$$SCON = \sum_L L^2 P^D(L)$$

Angular Second Moment:

$$\sum_K \left[P^S(K)\right]^2 \sum_L \left[P^D(L)\right]^2$$

Correlation:

$$COR = \frac{1/2 \left[\sum_K \left(K - \mu_S\right)^2 P^S(K) - \sum_L L^2 P^D(L) \right]}{\sigma^2}$$

Entropy:

$$SENT = -\sum_K P^S(K) \log P^S(K) - \sum_L P^D(L) \log P^D(L)$$

Local Homogeneity:

$$HOM = \sum_L \frac{P^D(L)}{\left(1 + L^2\right)}$$

Cluster Shade for Sum:

$$SSHAD = \sum_K \frac{\left(K - \mu_S\right)^3 P^S(K)}{\sigma^3}$$

Cluster Prominence for Sum:

$$SPROM = \frac{\sum_K \left(K - \mu_S\right)^4 P^S(K)}{\sigma^4} - 3$$

Appendix 3
Gray Level Difference Vector Features [Ref. 11]

From the computed probability density function $P_d(m)$ of the absolute gray level difference of two pixels separated by a vector distance d one can compute the following features:

Mean:

$$\mu_d = \sum_m m P_d(m)$$

Standard Deviation:

$$\sigma_d = \left[\sum_m (m - \mu_d)^2 P_d(m) \right]^{1/2}$$

Contrast:

$$CON_d = \sum_m m^2 P_d(m)$$

Angular Second Moment:

$$ASM_d = \sum_m [P_d(m)]^2$$

Entropy:

$$ENT_d = -\sum_m P_d(m) \log P_d(m)$$

Local Homogeneity:

$$HOM_d = \sum_m \frac{P_d(m)}{(1 + m^2)}$$

Cluster Shade:

$$SHAD_d = \frac{\left[\sum_m (m - \mu_d)^3 P_d(m) \right]}{\sigma_d^{\ 3}}$$

Cluster Prominence:

$$PROM_d = \frac{\left[\sum_m (m - \mu_d)^4 P_d(m) \right]}{\sigma_d^{\ 4}} - 3$$